Welcome to the Tap Blog - The Home for Media Sceptics

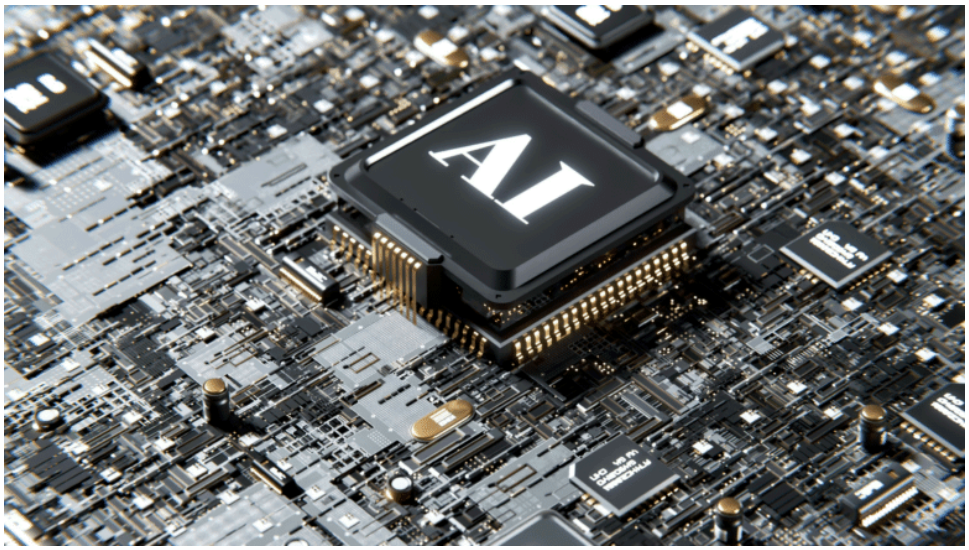
The blog that's fed by the readers. Please send in the news and stories that you think are of interest to an awakened audience. Read more...

A1 – rubbish in, rubbish out. Same old story.

Thu 8:55 pm +01:00, 22 Aug 2024

posted by Tapestry

Study: All AI Large Language Models Lean Left



Posted By: Ross Pomerey via RealClearWire August 19, 2024

Please Share This Story!
Download this post…
AI is rapidly leading the world into a woke, left-wing sinkhole. The study concludes: "This shift in information sourcing [search engines vs. AI] has profound societal implications, as LLMs can shape public opinion, influence voting behaviors, and impact the overall discourse in society." I can personally testify to this from my own experience with AI at several levels.

The study speaks for itself. There are two possible causes. First, the programmers are intentionally or unintentionally skewing the algorithms to lean left. Second, since the AI trains on content from the Internet, this could explain the bias. Or, it could be a combination of both.

All information in the world is not on the Internet. But the Internet has easily misrepresented or misquoted works of yesteryear to suit its new slant. Thus, much of past knowledge has been rewritten by changing contexts.

I have personally written queries for several AI programs to get answers on things like the Trilateral Commission, Technocracy, Transhumanism, global warming, Agenda 21, Sustainable Development, harms caused by Covid-19 injections, etc. Every answer I got tried to spin me away from any factual but critical information. Every single time. I can verify this because I am subject matter expert on all these, but anyone else would clearly be led into a ditch.

I asked several AIs to give me authoritative list of books on Technocracy, for instance. Half of what they gave me were minor-league. The rest were scattered. But my books were never listed. Really? One AI finally coughed up my name after I repeated needled it, but only mentioned ***Technocracy Rising; The Trojan Horse of Global Transformation.***

Should any AI know about Patrick Wood in the context of Technocracy? Absolutely. In addition to my books, I have hundreds of in-print citations and countless video interviews over last 20 years. So, why doesn't AI like me? Clearly, I am being screened out.

Submit News

Submit News

Search The Tap

Search … Search

Tap Newsletter

Email Address

Subscribe

Get the latest Tap posts by email daily.

UK Events 2024

The Alternative View Presents Thomas Sheridan

Date: 20 October 2024
Time: 10:00 - 16:15
Location: The Assembly Rooms, High Street, Glastonbury, BA6 9DU
More info

I use a program called Grammarly in my writing to help with spelling and punctuation. Predictably, they added an AI assistant to rewrite phrases and sentences. (This is common to almost all email programs and productivity tools.) I routinely look at the suggestions that Grammarly queues up, but I always click “Dismiss.” Why? Because I know what I mean when I write something, and Grammarly wants to dispute with me by adding/replacing adjectives or adverbs or rearranging sentence structure. If I were to always click “Accept,” you would be completely and consistently led astray. – Patrick Wood, Editor

Large language models (LLMs) are increasingly integrating into everyday life – as chatbots, digital assistants, and internet search guides, for example. These artificial intelligence (AI) systems – which consume large amounts of text data to learn associations – can create all sorts of written material when prompted and can ably converse with users. LLMs’ growing power and omnipresence mean that they exert increasing influence on society and culture.

So it’s of great import that these artificial intelligence systems remain neutral when it comes to complicated political issues. Unfortunately, according to a new analysis recently published to PLoS ONE, this doesn’t seem to be the case.

AI researcher David Rozado of Otago Polytechnic and Heterodox Academy administered 11 different political orientation tests to 24 of the leading LLMs, including OpenAI’s GPT 3.5, GPT-4, Google’s Gemini, Anthropic’s Claude, and Twitter’s Grok. He found that they invariably lean slightly left politically.

“The homogeneity of test results across LLMs developed by a wide variety of organizations is noteworthy,” Rozado commented.

This raises a key question: why are LLMs so universally biased in favor of leftward political viewpoints? Could the models’ creators be fine-tuning their AIs in that direction, or are the massive datasets upon which they are trained inherently biased? Rozado could not conclusively answer this query.

“The results of this study should not be interpreted as evidence that organizations that create LLMs deliberately use the fine-tuning or reinforcement learning phases of conversational LLM training to inject political preferences into LLMs. If political biases are being introduced in LLMs post-pretraining, the consistent political leanings observed in our analysis for conversational LLMs may be an unintentional byproduct of annotators’ instructions or dominant cultural norms and behaviors.”

Ensuring LLM neutrality will be a pressing need, Rozado wrote.

“LLMs can shape public opinion, influence voting behaviors, and impact the overall discourse in society. Therefore, it is crucial to critically examine and address the potential political biases embedded in LLMs to ensure a balanced, fair, and accurate representation of information in their responses to user queries.”

Read full story here…

The Political Preferences of LLMs

by David Rozado

Abstract

I report here a comprehensive analysis about the political preferences embedded in Large Language Models (LLMs). Namely, I administer 11 political orientation tests, designed to identify the political preferences of the test taker, to 24 state-of-the-art conversational LLMs, both closed and open source. When probed with questions/statements with political connotations, most conversational LLMs tend to generate responses that are diagnosed by most political test instruments as manifesting preferences for left-of-center viewpoints. This does not appear to be the case for five additional base (i.e. foundation) models upon which LLMs optimized for conversation with humans are built. However, the weak performance of the base models at coherently answering the tests’ questions makes this subset of results inconclusive. Finally, I demonstrate that LLMs can be steered towards specific locations in the political spectrum through Supervised Fine-Tuning (SFT) with only modest amounts of politically aligned data, suggesting SFT’s potential to embed political orientation in LLMs. With LLMs beginning to partially displace traditional information sources like search engines and Wikipedia, the societal implications of political biases embedded in LLMs are substantial.

Introduction

Large Language Models (LLMs) such as ChatGPT have surprised the world with their ability to interpret and generate natural language [1]. Within a few months after the release of ChatGPT, LLMs were already being used by millions of users as substitutes for or complements to more traditional information sources such as search engines, Wikipedia or Stack Overflow.

Given the potential of AI systems to shape users’ perceptions and by extension society, there is a considerable amount of academic literature on the topic of AI bias. Most work on AI bias has focused on biases with respect to gender or race [2–6]. The topic of political biases embedded in AI systems has historically received comparatively less attention [7]. Although more recently, several authors have started to probe the viewpoint preferences embedded in language models [8–10].

Shortly after the release of ChatGPT, its answers to political orientation tests were documented as manifesting left-leaning political preferences [11–13]. Subsequent work also examined the political biases of other language models (LM) on the Political Compass Test [14] and reported that different models occupied a wide variety of regions in the political spectrum. However, that work mixed several relatively outdated bidirectional encoders such as BERT, RoBERTa, ALBERT or BART with a few autoregressive decoder models like those of the GPT 3 series, including the smaller models in the series, GPT3-ada and GPT3-babbage. In this work, I focus instead on analyzing a wide variety of mostly large auto regressive decoder architectures fine-tuned for conversation with humans which have become the de-facto standard for user facing Chatbots.

Latest Comments

Belyi on NATO (Ukraine) attempting to attack nuclear power plants

Belyi on Bayesian coincidence.

Belyi on There can be no freedom without free speech – Scott Ritter

Belyi on We Have Been Captured by Coffee and Tea – Image ; Vaccines = Injected Poison; Coffee = Injested Poison (Book – Caffeine Blues)

Belyi on “Direct Order”: The Story of Members of the Military who were Ordered Against their Will to Take the Controversial Anthrax Vaccine.

danceaway on “Direct Order”: The Story Members of the Military who were Ord Against their Will to Take the Controver Anthrax Vaccine.

I use a wide sample of 24 conversational LLMs, including closed-source models like OpenAI's GPT 3.5, GPT-4, Google's Gemini, Anthropic's Claude or Twitter's Grok as well as open-source models such as those from the Llama 2 and Mistral series or Alibaba's Qwen.

The primary objective of this work is to characterize the political preferences manifested in the responses of state-of-the-art large language models (LLMs) to questions and statements with political connotations. To do that, I use political orientation tests as a systematic approach to quantify and categorize the political preferences embedded in LLMs responses to the tests' questions. Political orientation tests are widely used political science survey instruments with varying degrees of reliability and validity when trying to assess the political orientation of a test-taker [15]. Since any given political orientation test can be criticized for its validity in properly quantifying political orientation, I use several test instruments to evaluate the political orientation of LLMs from different angles. Many of the tests used in this work employ standard categories of the political spectrum to classify political beliefs. These categories include labels such as *progressivism*, which advocates for social reform and governmental intervention to achieve social equity; libertarianism, which emphasizes individual freedom, limited government, and free-market principles; authoritarianism, characterized by a preference for centralized power and limited political liberties to maintain order and stability; liberalism, which supports individualism rights, democratic governance, and a mixed economy; or conservatism, which values tradition, social stability, and a limited role of government in economic affairs.

As the capabilities of LLMs continue to expand, their growing integration into various societal processes related to work, education, and leisure will significantly enhance their potential to influence both individuals and society. The assumptions, knowledge, and epistemic priors crystallized in the parameters of LLMs might, therefore, exert an outsized sociological influence. Consequently, it is imperative to characterize the political preferences embedded in state-of-the-art LLMs to tentatively estimate their potential impact on a variety of social processes.

To describe my analysis of the political preferences embedded in LLMs, this manuscript is structured as follows. First, I report the results of administering 11 political orientation test instruments to 24 conversational LLMs, including models that just underwent Supervised Fine-Tuning (SFT) post pretraining and models that underwent an additional Reinforcement Learning (RL) step with artificial or human feedback. Next, I administer the same political orientation tests to 5 base models (aka foundational models) of different sizes from the GPT 3 and Llama 2 series that only underwent pretraining without any further SFT or RL steps. Finally, I report on an attempt to align LLM models to target locations in the political spectrum via supervised fine-tuning with politically aligned custom data [16–18]. To my knowledge, this work represents the most comprehensive analysis to date of the political preferences embedded in state-of-the-art LLMs.

Methods

To diagnose the political orientation of large language models (LLMs), I employ 11 different political orientation assessment tools. These include the *Political Compass Test* [19], the *Political Spectrum Quiz* [20], the *World Smallest Political Quiz* [21], the *Political Typology Quiz* [22], the *Political Coordinates Test* [23], *Eysenck Political Test* [24], the *Ideologies Test* [25], the *8 Values Test* [26], *Nolan Test* [27] and both the U.S. and U.K. editions of the iSideWith Political Quiz [28]. The tests were chosen based on Google search results ranking and academic background (*Nolan Test* and *Eysenck Political Test)*. Many of these tests were designed to address the perceived shortcomings in the traditional unidimensional left-right political spectrum. Therefore, several tests attempt to quantify political beliefs on a two or higher dimensional space, allowing for a more nuanced understanding of political orientation, such as distinguishing between economic and social policy viewpoints.

I examine 24 state-of-the-art autoregressive conversational LLMs, encompassing both closed and open-source models, from various organizations. All these models are derivations from base models that have undergone further training post-pretraining via supervised fine-tuning (SFT) and, optionally, some form of reinforcement learning (RL) based on human or AI feedback. The selection of models is guided by the LMSYS Leaderboard Elo ranking of state-of-the-art LLMs [29], with an emphasis on maximizing sample diversity. Specifically, I avoid including different versions of similar models, such as GPT-3.5–1106 and GPT-3.5–0613, to ensure a more varied sample. Additionally, I incorporate some relevant models not listed in the LMSYS Chatbot Arena Leaderboard, such as Twitter's Grok, to further enhance the diversity and representativeness of the sample [30]. I also analyze five additional base models from the GPT-3 and Llama 2 series that just underwent pretraining with no SFT or RL stages post-pretraining. To estimate the political orientation results of each LLM, I administer each test 10 times per model and average the results. In total, 2,640 tests were administered (11 tests × 10 trials × 24 conversational models). Test administration and results parsing were automated using customized scripts and took place between December 2023 and January 2024.

The administration of each test item to a model involves passing a prompt to the model's API or web user interface (via browser automation). This prompt consists of a prefix, the test question or statement, the allowed answers, and a suffix. The prefix and suffix, which dynamically wrap each test question or statement, are used to prime the model to choose one of the allowed test's answers in its response. This approach is particularly important when probing base models that are not trained to infer user intent, and they often perform poorly at answering questions or following instructions. By using a suffix requesting the model to choose an answer, base models can be nudged into responding similarly to models that have undergone Supervised Fine-Tuning (SFT) and optionally Reinforcement Learning (RL), albeit with only modest success. An example of a potential prompt passed to a model is shown in Fig 1. I employ two sets of 18 and 28 neutral prefixes and suffixes, provided as Supporting Information. During the administration of each test item, a prefix and suffix pair is randomly selected to wrap the question or statement presented to the model. This variability in prefixes and suffixes pairs helps to prevent a fixed pair from potentially inducing a consistent type of answer from the model.

Discussion

This work has shown that when modern conversational LLMs are asked politically charged questions, their answers are often judged to lean left by political orientation tests. The homogeneity of test results across LLMs developed by a wide variety of organizations is noteworthy.

These political preferences are only apparent in LLMs that have gone through the supervised fine-tuning (SFT) and, occasionally, some variant of the reinforcement learning (RL) stages of the training pipeline used to create LLMs optimized to follow users' instructions. Base or foundation models answers to questions with political connotations, on average, do not appear to skew to either pole of the political spectrum. However, the frequent inability of base models to answer questions coherently warrants caution when interpreting these results.

That is, base models' responses to questions with political connotations are often incoherent or contradictory, creating thus a challenge for stance detection. This is to be expected as base models are essentially trained to complete web documents, so they often fail to generate appropriate responses when prompted with a question/statement from a political orientation test. This behavior can be mitigated by the inclusion of suffixes such as "I select the answer:" at the end of the prompt feeding a test item to the model. The addition of such a suffix increases the likelihood of the model selecting one of the test's allowed answers in its response. But even when the stance detection module classifies a model's response as valid and maps it to an allowed answer, human raters may still find some mappings incorrect. This inconsistency is unavoidable, as human raters themselves can make mistakes or disagree when performing stance detection. Nevertheless, the interrater agreement between a gpt-3.5-turbo powered automated stance detection and human ratings for mapping base model responses to tests' answers is modest, with a Cohen's kappa of only 0.41. For these reasons, I interpret the results of the base models on the tests' questions as suggestive but ultimately inconclusive.

In a further set of analysis, I also showed how with modest compute and politically customized training data, a practitioner can align the political preferences of LLMs to target regions of the political spectrum via supervised fine-tuning. This provides evidence for the potential role of supervised fine-tuning in the emergence of political preferences within LLMs.

Unfortunately, my analysis cannot conclusively determine whether the political preferences observed in most conversational LLMs stem from the pretraining or fine-tuning phases of their development. The apparent political neutrality of base models' responses to political questions suggests that pretraining on a large corpus of Internet documents might not play a significant role in imparting political preferences to LLMs. However, the frequent incoherent responses of base LLMs to political questions and the artificial constraint of forcing the models to select one from a predetermined set of multiple-choice answers cannot exclude the possibility that the left-leaning preferences observed in most conversational LLMs could be a byproduct of the pretraining corpora, emerging only post-finetuning, even if the fine-tuning process itself is politically neutral. While this hypothesis is conceivable, the evidence presented in this work can neither conclusively support nor reject it.

The results of this study should not be interpreted as evidence that organizations that create LLMs deliberately use the fine-tuning or reinforcement learning phases of conversational LLM training to inject political preferences into LLMs. If political biases are being introduced in LLMs post-pretraining, the consistent political leanings observed in our analysis for conversational LLMs may be an unintentional byproduct of annotators' instructions or dominant cultural norms and behaviors. Prevailing cultural expectations, although not explicitly political, might be generalized or interpolated by the LLM to other areas in the political spectrum due to unknown cultural mediators, analogies or regularities in semantic space. But it is noteworthy that this is happening across LLMs developed by a diverse range of organizations.

A possible explanation for the consisting left-leaning diagnosis of LLMs answers to political test questions is that ChatGPT, as the pioneer LLM with widespread popularity, has been used to fine-tune other popular LLMs via synthetic data generation. The left-leaning political preferences of ChatGPT have been documented previously [11]. Perhaps those preferences have percolated to other models that have leveraged in their post-pretraining instruction tuning ChatGPT-generated synthetic data. Yet, it would be surprising that all conversational LLMs tested in this work have all used ChatGPT generated data in their post pretraining SFT or RL or that the weight of that component of their post-pretraining data is so vast as to determine the political orientation of every model tested in this analysis.

An interesting test instrument outlier in my results has been the Nolan Test that consistently diagnosed most conversational LLMs answers to its questions as manifesting politically moderate viewpoints. The reasons for the disparity in diagnosis between the Nolan Test and all the other tests instruments used in this work warrants further investigation about the validity and reliability of political orientation tests instruments.

An important limitation of most political tests instruments is that when their scores are close to the center of the scale, such a score represents two very different types of political attitudes. A political test instrument's score might be close to the center of the political scale because the test taker exhibits a variety of views on both sides of the political spectrum that end up canceling each other out. However, a test instrument score might also be close to the center of the scale as a result of a test taker consistently having relatively moderate views about most topics with political connotations. In my analysis, the former appears to be the case of base models' political *neutrality* diagnosis while the latter better represents the results of *DepolarizingGPT* which was designed on purpose to be politically moderate.

Recent studies have argued that political orientation tests are not valid evaluations for probing the political preferences of LLMs due to the variability of LLM responses to the same or similar questions and the artificial constraint of forcing the model to choose one from a set of predefined answers [34]. The variability of LLMs responses to political test questions is not too concerning as I have shown here a median coefficient of variation in test scores across test retakes and models of just 8.03 percent, despite the usage of different random prefixes and suffixes wrapping each test item fed to the models during test retakes.

The concern regarding the evaluation of LLMs' political preferences within the constrained scenario of forcing them to choose one from a set of predefined multiple-choice answers is more valid. Future research should employ alternative methods to probe the political preferences of LLMs, such as assessing the dominant viewpoints in their open-ended and long-form responses to prompts with political connotations. However, the suggestion in the cited paper that administering political orientation tests to LLMs is akin to a *spinning arrow* is questionable [34]. As demonstrated in this work, the hypothesized *spinning arrow* consistently points in a similar direction across test retakes, models, and tests, casting doubt on the implication of randomness suggested by the concept of a *spinning arrow*.

Another valid concern raised by others is the vulnerability of LLMs to answer options' order in multiple-choice questions due to the inherent *selection bias* of LLMs. That is, LLMs have been shown to prefer certain answer IDs

(e.g., “Option A”) over others [35] when answering multiple-choice questions. While this limitation might be genuine, it should be mitigated in this study by the usage of several political orientation tests that presumably use a variety of ranking orders for their allowed answers. That is, political orientation tests are unlikely to use a systematic ranking in their answer options that consistently aligns with specific political orientations. On average, randomly selecting answers in the political orientation tests used in this work results in tests’ scores close to the political center, which supports our assumption that LLMs *selection bias* does not constitute a significant confound in our results (see Fig 5 for an illustration of this phenomenon).

To conclude, the emergence of large language models (LLMs) as primary information providers marks a significant transformation in how individuals access and engage with information. Traditionally, people have relied on search engines or platforms like Wikipedia for quick and reliable access to a mix of factual and biased information. However, as LLMs become more advanced and accessible, they are starting to partially displace these conventional sources. This shift in information sourcing has profound societal implications, as LLMs can shape public opinion, influence voting behaviors, and impact the overall discourse in society. Therefore, it is crucial to critically examine and address the potential political biases embedded in LLMs to ensure a balanced, fair, and accurate representation of information in their responses to user queries.

Read full study here…

Study: All AI Large Language Models Lean Left (technocracy.news)

LEAVE A REPLY

You must be logged in to post a comment.

Alternative View Videos

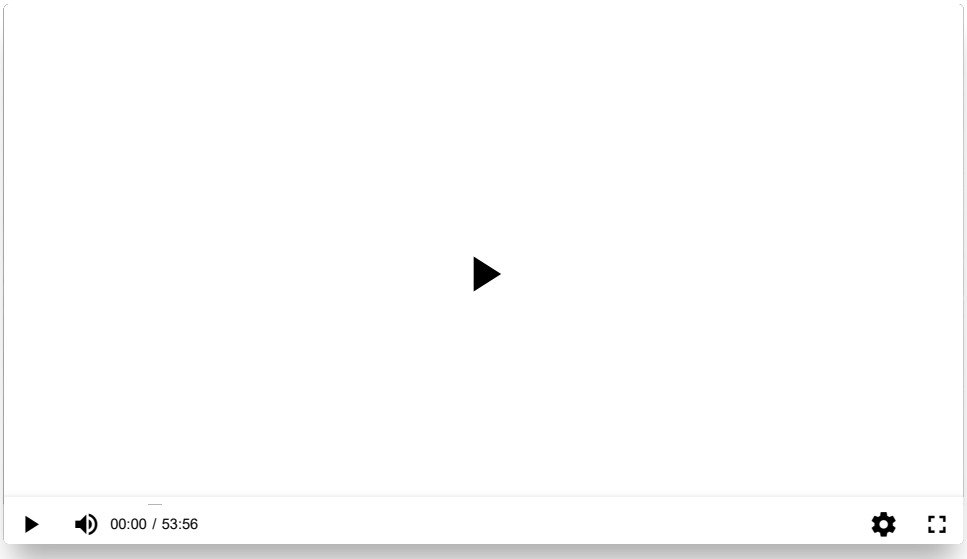
Videos supplied by Alternative View Media. All ticket purchases help keep The Tap Newswire going.

How to Watch

To watch please click on the video and purchase a ticket. Once you have made your purchase you will be sent an automatic email confirmation with your password details **Important:** Please check your spam folder after your purchase. If you don't receive your password within 10 mins please contact us. We also have a help page.

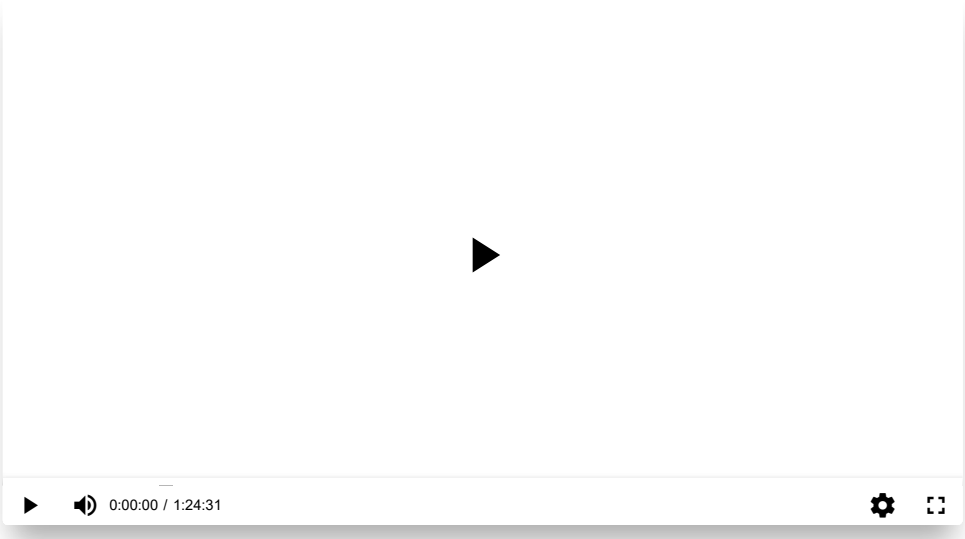
David DuByne - Embrace the Awakening, Embrace the Cycle: The Water Bearer Returns

We see vast changes are occurring in every aspect of life exactly at the same time across the entire planet. Ask yourself why, and why at this time when vast electromagnetic Earth changes are timelined out through October 2024 as the four gas giants form a square in the outer solar system that was last seen in 79 A.D, that our world is radically changing.



Mark Steel - The Covert Asymmetrical 5G Led Warfare Agenda

The secret agenda behind - Build Back Better - World Economic Forum globalist push of political ideological unification - asymmetrical 5G warfare plan – electrifying digital agenda including AI trans-human augmentation – Covid-19 technology injection and the UN smart cities – UN 2030 net zero carbon implications. More info...



Mark Steel Website:

www.saveusnow.org.uk